



Residential and Commercial Building Classification Based on Street-Level Images Using Deep Learning

**Najmi Fadhila Atsari¹, Tisha Melia^{2*},
Aland Polma Naek Sihombing³, Fatayat⁴**

^{1,3,4}Department of Computer Science, Faculty of Mathematics and Natural Sciences,
Universitas Riau, Indonesia

²Department of Information System, Faculty of Science and Technology,
Universitas Prima Indonesia, Indonesia

E-Mail: ¹najmi.fadhila0233@student.unri.ac.id, ²tisha.melia@lecturer.unri.ac.id,

³alandfolmanaeksihombing@unprimdn.ac.id, ⁴fatayat@lecturer.unri.ac.id

Received Apr 13th 2026; Revised Jun 07th 2026; Accepted Jul 23th 2026; Available Online Jun 31th 2026

Corresponding Author: Tisha Melia

Copyright © 2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

The classification of buildings into residential and commercial categories is essential for computer vision applications that support urban analysis and automated decision-making. Accurate identification of building functions from visual data enables downstream applications, such as service differentiation and resource allocation. However, manual image-based classification is inefficient and subjective, particularly when applied to large-scale datasets. This study proposes a deep learning approach to classifying residential and commercial buildings using street-level images. A dataset comprising 882 images was collected from publicly available sources and categorized based on visual characteristics. The dataset was divided into training, validation, and test sets in a 60:20:20 ratio. The preprocessing stages included data cleaning, labeling, cropping, and resizing. Two models were evaluated: a conventional Convolutional Neural Network (CNN) and MobileNetV2 with transfer learning. Model performance was optimized by tuning the batch size and learning rate. The experimental results show that MobileNetV2 achieved the best performance with a batch size of 32 and a learning rate of 0.0001, attaining an accuracy of 92.05% and precision, recall, and F1-score values of 91.46%. Evaluation using a confusion matrix indicated low misclassification rates for both building categories. These results demonstrate that deep learning models can effectively classify buildings based on visual data.

Keywords: Building Classification, CNN, Computer Vision, Google Street Images, MobileNetV2.

1. INTRODUCTION

The rapid expansion of urban environments has intensified the need for automated, scalable methods to assess the functional characteristics of buildings, particularly distinguishing residential from commercial uses [1]. Accurate identification of building functions supports a wide range of applications, including urban planning, infrastructure provision, smart city services, disaster response, and socioeconomic analysis [2], [3], [4]. Recent advances in computer vision and deep learning have enabled significant progress in extracting semantic information from visual data, making image-based building analysis a practical alternative to labor-intensive field surveys and manual annotation processes [5], [6], [7].

Early research in this domain predominantly focused on building detection and extraction from high-resolution remote sensing imagery [8]. Convolutional Neural Networks (CNNs) and fully convolutional architectures have demonstrated strong performance in delineating building footprints under complex urban conditions [9], [10]. To address challenges such as scale variation, occlusion, and heterogeneous urban morphology, subsequent studies incorporated multiscale feature fusion, attention mechanisms, and cross-level semantic integration [11], [12], [13]. Transformer-based architectures and hybrid Convolutional Neural Network (CNN)-Transformer architectures were subsequently introduced to enhance global contextual awareness, thereby further improving the accuracy of building extraction and segmentation across diverse datasets [14], [15], [16]. Although these approaches achieved impressive results in identifying building locations, their primary objective remained spatial extraction rather than functional interpretation.

Recognizing the limitations of purely geometric analysis, recent studies have shifted toward classifying building types and uses, emphasizing the importance of functional attributes in urban analytics. Building type

classification has been investigated using very-high-resolution satellite and aerial imagery, with Convolutional Neural Network (CNN)-based and hybrid Convolutional Neural Network (CNN)-Transformer models demonstrating the ability to distinguish residential structures from nonresidential structures with high accuracy [17]. However, these methods often face challenges related to interclass similarity and class imbalance, particularly in dense urban areas where residential and commercial buildings may share similar rooftop structures and spatial layouts [18]. Furthermore, most existing studies on building type classification rely on overhead imagery, thereby limiting the availability of street-level visual cues essential for fine-grained functional differentiation.

In contrast, street-level imagery provides a richer representation of building appearances by capturing façade characteristics, entrances, signage, window layouts, and architectural details that often indicate building use. Studies on scene recognition and architectural image analysis have demonstrated that deep learning models can effectively classify built environments using ground-level images [19]. Large-scale analyses of street-level imagery have further revealed both the strengths and limitations of deep learning systems, including socioeconomic and contextual biases that may influence classification outcomes [20], [21]. These findings highlight the importance of carefully designed datasets and models when applying deep learning to real-world urban imagery.

In addition to model expressiveness, computational efficiency has emerged as a critical consideration for practical deployment. Lightweight Convolutional Neural Network (CNN) architectures, such as MobileNet variants, are specifically designed to reduce parameter count and computational cost while maintaining competitive classification performance [22]. Previous studies have shown that transfer learning with lightweight models can achieve high accuracy using limited datasets, making these models suitable for real-world applications and edge deployment scenarios [15]. Comparative analyses of conventional Convolutional Neural Networks (CNNs) and lightweight architectures indicate that efficiency-oriented models often provide a favorable trade-off between accuracy and resource consumption [9], [19].

Despite these advances, important research gaps remain. Existing studies on building extraction primarily focus on identifying building footprints and spatial boundaries from aerial or remote-sensing imagery rather than on determining building functions [6], [23], [24]. Meanwhile, studies using street-level imagery generally address scene recognition, architectural analysis, or urban perception without explicitly distinguishing between residential and commercial building uses [17], [19], [25], [26], [27]. Furthermore, although lightweight architectures such as MobileNet have demonstrated strong performance in image classification tasks, their application and comparison with conventional Convolutional Neural Network (CNN) models for street-level building function classification remain limited [15], [22]. Therefore, further investigation is needed to evaluate efficient deep learning approaches for the practical classification of building uses in urban environments.

To address these gaps, this study proposes a deep learning-based framework for classifying residential and commercial buildings using street-level imagery. Two architectures, namely a conventional Convolutional Neural Network (CNN) and MobileNet Version 2 (MobileNetV2) with transfer learning, are systematically compared. In addition to evaluating classification performance, this study investigates the effects of key hyperparameters and assesses the suitability of lightweight deep learning models for practical urban applications. The findings contribute to urban computer vision research and provide insights into applications such as service differentiation, small-business identification, and urban digital infrastructure planning.

2. MATERIALS AND METHOD

This study employed an experimental research design to evaluate deep learning models for classifying residential and commercial buildings using street-level images. The overall workflow consisted of five main stages: (1) data acquisition, (2) data preprocessing, (3) model construction, (4) model training and hyperparameter tuning, and (5) performance evaluation. Two deep learning architectures were implemented and compared, namely a conventional Convolutional Neural Network (CNN) and the MobileNetV2 architecture with transfer learning.

2.1. Data Acquisition

A total of 882 street-level building images were used in this study (Fig.1). The images were collected from Google Street View and Google Images, which provide diverse real-world representations of buildings captured under varying lighting conditions, viewing angles, and urban contexts. The final dataset consisted of 441 residential building images and 441 commercial building images.

Each image was manually annotated into one of two classes based on observable visual characteristics:

1. Residential buildings, characterized by the absence of commercial signage, storefront displays, or advertising elements.
2. Commercial buildings, characterized by the presence of visual indicators such as shop signs, banners, display windows, or other commercial attributes.

To ensure labeling consistency, all images were reviewed and annotated using predefined classification criteria based on visible façade characteristics and building usage indicators. Images with severe occlusion, extremely low resolution, multiple dominant buildings, or ambiguous building functions were excluded from the dataset. The collected images represent a variety of building appearances and environmental conditions, including differences in façade design, lighting, weather conditions, and viewing perspectives. Such variability was intentionally preserved to improve the robustness of the classification models under realistic urban scenarios. In illustrated Figure 1. The complete dataset has been publicly released and is available at <https://my.unri.ac.id/keseluruhandata>.

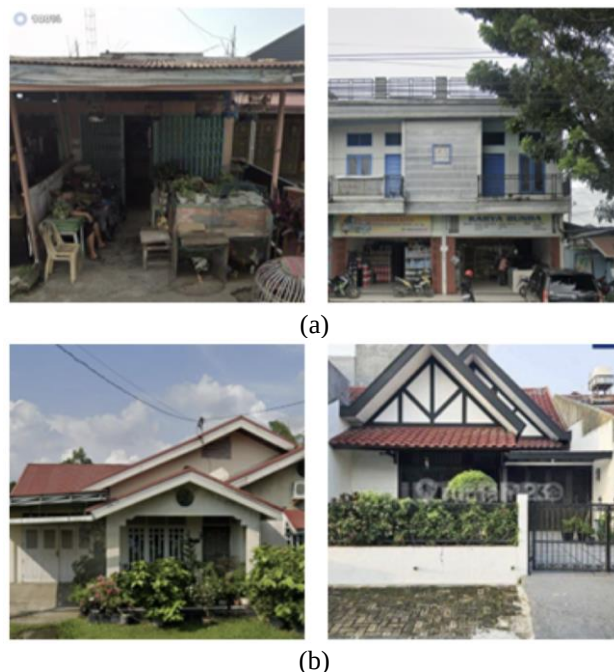


Figure 1. Sample images from the dataset illustrating the two building categories used in this study. (A) Commercial buildings with storefronts and business signage. (B) Residential buildings representing typical housing façades without visible commercial activity.

2.2. Data Preprocessing

All images were resized to 224×224 pixels, a standard input size for Convolutional Neural Network (CNN)-based architectures and one required by the MobileNetV2 model. Image cropping was applied when necessary to focus on the main building region and reduce background interference. The pixel values were normalized to the range of $[0, 1]$ to stabilize the training process. The dataset was divided into training, validation, and testing subsets, with proportions of 60%, 20%, and 20%, respectively. The training set was used to learn the model parameters, the validation set was used for hyperparameter tuning, and the testing set was reserved exclusively for the final performance evaluation.

2.3. Model Construction

There are two models that we build and compare in this study:

1. Convolutional Neural Network (CNN)

The baseline CNN model consisted of sequential convolutional layers using Rectified Linear Unit (ReLU) activation functions, followed by max-pooling layers for spatial downsampling. Dropout layers were incorporated to reduce overfitting. The extracted feature maps were then flattened and passed through fully connected layers. The final layer used a sigmoid activation function for binary classification.

2. MobileNetV2 with Transfer Learning

The second model employed the MobileNetV2 architecture, which uses depthwise separable convolutions and inverted residual blocks to reduce computational complexity. Pretrained weights obtained from a large-scale image dataset were used to initialize the network. The original classification layer was removed and replaced with a custom fully connected layer, followed by a sigmoid activation function tailored to binary classification. During the initial training stage, the convolutional base was

frozen to preserve the previously learned feature representations. Selected layers were subsequently fine-tuned to improve classification performance.

2.4. Model Training and Hyperparameter Settings

Model training was conducted using the TensorFlow and Keras frameworks. The Adam optimizer was employed with different learning rates to identify the optimal configuration. Learning rates of 0.01, 0.001, 0.0001, and 0.00001 were evaluated to represent a range from relatively aggressive to conservative optimization settings commonly used in transfer learning applications. Batch sizes of 32, 64, and 128 were also investigated to assess the trade-off between training stability and computational efficiency. These values were selected based on commonly adopted settings in image classification studies and preliminary experiments that indicated stable model convergence. The models were trained for up to 50 epochs, with early stopping applied based on validation loss to prevent overfitting. Binary cross-entropy was used as the loss function for both models.

2.5. Performance Evaluation

Model performance was assessed using multiple evaluation metrics, including accuracy, precision, recall, and F1-score. A confusion matrix was employed to analyze classification outcomes for residential and commercial buildings in terms of true positives, true negatives, false positives, and false negatives. These metrics provided a comprehensive evaluation of classification effectiveness and class-level prediction behavior.

3. RESULTS AND DISCUSSION

The training performance of the proposed models was evaluated to understand how effectively each architecture learned discriminative visual features from the street-level building dataset. Two deep learning models were implemented and compared, namely a baseline CNN and MobileNetV2 with transfer learning. The evaluation focused on training behavior across different batch size configurations as well as convergence patterns during the training process.

Table 1 and Figure 2 summarize the training and validation performance of CNN and MobileNetV2 across different batch-size configurations. MobileNetV2 consistently outperformed the baseline CNN across all evaluated metrics, demonstrating superior capability in learning discriminative features from street-level building images. The highest validation accuracy was achieved by MobileNetV2 with a batch size of 64 (95.25%), compared with the best CNN validation accuracy of 86.36%. Similarly, MobileNetV2 achieved substantially higher training accuracy, precision, and recall values across all configurations. For example, at a batch size of 32, MobileNetV2 achieved training and validation accuracies of 97.82% and 94.38%, respectively, whereas CNN achieved 88.06% and 86.30%. These findings suggest that the combination of transfer learning and the efficient architecture of MobileNetV2 enables more effective feature extraction and improved generalization compared with a conventional CNN.

Table 1. Training and Validation Performance of MobileNetV2 and CNN Models Under Different Batch Size Configurations.

Model/ Batch Size	MobileNetV2			CNN		
	32	64	128	32	64	128
Training Accuracy	0,9782	0,9726	0,9551	0,8806	0,8612	0,8436
Validation Accuracy	0,9438	0,9525	0,9466	0,8630	0,8636	0,8466
Training Loss	0,0672	0,0715	0,1163	0,2785	0,3194	0,3658
Validation Loss	0,1438	0,1345	0,1421	0,3520	0,3598	0,4023
Recall	0,9799	0,9759	0,9558	0,8902	0,8746	0,8559
Precision	0,9789	0,9739	0,9586	0,8824	0,8734	0,8487

The superior performance of MobileNetV2 can be attributed to both its architectural design and the use of transfer learning. Unlike conventional CNNs, MobileNetV2 employs depthwise separable convolutions and inverted residual blocks, which enable efficient feature extraction while reducing the number of trainable parameters. This architecture enables the model to capture relevant visual patterns in building façades without introducing excessive complexity. Furthermore, MobileNetV2 was initialized with pretrained weights learned on large-scale image datasets, providing robust low- and high-level visual representations prior to training on the target dataset. In contrast, the conventional CNN was trained from scratch and therefore had to learn all feature representations directly from the available training data. These advantages likely contributed to the higher accuracy, precision, and recall achieved by MobileNetV2 across all evaluated batch-size configurations in the illustrated Figure 2.

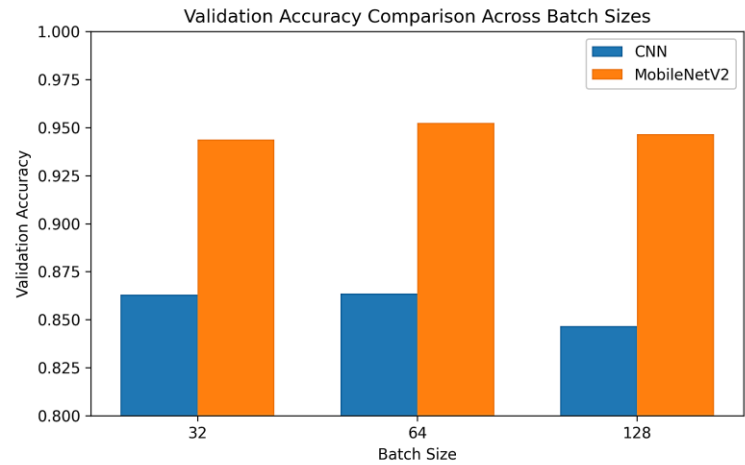


Figure 2. Validation Accuracy Comparison Between CNN and MobileNetV2 Across Different Batch-Size Configurations.

In addition to accuracy improvements, MobileNetV2 also produced substantially lower loss values during training. The training loss for MobileNetV2 at batch size 32 was recorded at 0.0672, which is considerably lower than the CNN loss of 0.2785 under the same configuration. Lower loss values indicate more effective parameter optimization and suggest that the MobileNetV2 architecture is better suited for extracting informative features from the dataset.

The influence of batch size on model performance can also be observed in Table 1. Increasing the batch size tended to reduce overall performance for both architectures. For instance, MobileNetV2 accuracy decreased from 0.9782 at batch size 32 to 0.9551 at batch size 128. A similar pattern was observed in the CNN model, where accuracy declined as batch size increased. Smaller batch sizes allow more frequent weight updates during training, which can help the model adapt more effectively to complex image features. As a result, batch size 32 was identified as the most suitable configuration for this dataset.

The training behavior of both models is further illustrated in Figure 3, which shows the training and validation accuracy curves across training epochs. The CNN model demonstrates gradual improvement during training; however, the validation accuracy fluctuates more noticeably compared to MobileNetV2. These fluctuations indicate that the CNN model has greater difficulty generalizing to unseen validation data. In contrast, the MobileNetV2 model shows a faster increase in accuracy during the early epochs and maintains relatively stable validation performance throughout the training process in illustrated Figure 3.

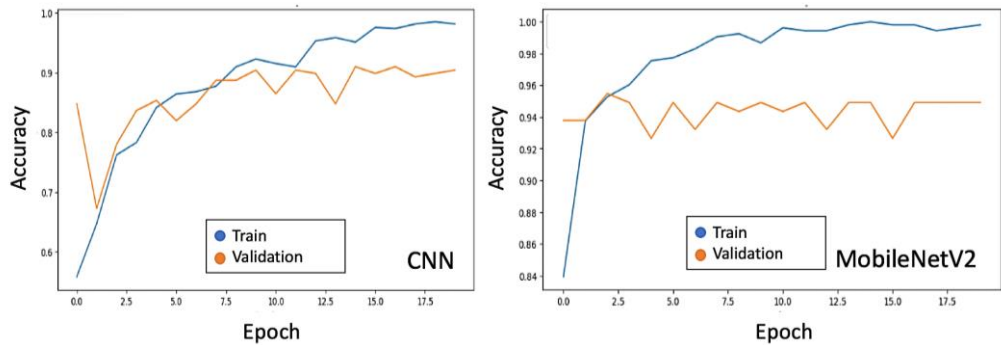


Figure 3. Training and Validation Accuracy Curves Across Epochs for the CNN (left) and MobileNetV2 (right) Models.

The smoother convergence pattern observed in MobileNetV2 can be attributed to the use of pretrained weights obtained from large-scale image datasets. Transfer learning allows the model to leverage previously learned visual representations, enabling faster convergence and improved generalization when applied to new tasks. This advantage is particularly important when the available dataset is moderate in size, as is the case in this study.

3.1 Effect of Learning Rate on Model Stability

Learning rate is a critical hyperparameter that influences the convergence speed and stability of neural network training. To determine the most appropriate learning rate for the MobileNetV2 model, several values

were evaluated, including 0.01, 0.001, 0.0001, and 0.00001. The resulting performance metrics are summarized in Table 2. The results show that a learning rate of 0.01 achieved the highest training accuracy (0.9807) and validation accuracy (0.9223). However, larger learning rates can cause unstable optimization behavior because the model parameters are updated with relatively large gradient steps. This can lead to overshooting the optimal solution and potentially degrading generalization performance, particularly when fine-tuning pretrained networks in illustrated Table 2.

Table 2. Training and Validation Performance of MobileNetV2 and CNN Models Under Different Learning Rates

Learning Rate	0,01	0,001	0,0001	0,00001
Training Accuracy	0,9807	0,9744	0,9756	0,9768
Validation Accuracy	0,9223	0,9169	0,9184	0,9198
Loss	0,0588	0,0695	0,0658	0,0623
Validation Loss	0,2196	0,2068	0,2295	0,2004
Recall	0,9804	0,9749	0,9749	0,9740
Precision	0,9803	0,9740	0,9760	0,9765

In contrast, extremely small learning rates such as 0.00001 produced stable loss values but required slower convergence, which can limit training efficiency. Although this configuration resulted in slightly lower training accuracy, the improvement in validation performance was not substantial enough to justify the increased training time. Among the tested configurations, a learning rate of 0.0001 provided the most balanced trade-off between training stability and classification performance. The model achieved a training accuracy of 0.9756 and a validation accuracy of 0.9184 while maintaining relatively consistent loss values throughout training. More importantly, this learning rate enabled gradual parameter updates, which is desirable when fine-tuning pretrained feature extractors such as MobileNetV2. Conservative learning rates help preserve previously learned visual representations while adapting the model to the target dataset. Based on these considerations, the learning rate of 0.0001 was selected as the optimal configuration for the final model because it provides stable convergence behavior and competitive classification performance without introducing excessive training instability.

3.3 Classification Performance on the Testing Set

Final evaluation was performed on the test set using the best-performing configuration, namely MobileNetV2 with a batch size of 32 and a learning rate of 0.0001. The confusion matrix in Figure 4 shows that most samples were correctly classified. The model achieved 75 true positives and 87 true negatives, with only 7 false positives and 7 false negatives. These values correspond to an overall accuracy of 92.05%, and precision, recall, and F1-score of approximately 91.46%. These results demonstrate that MobileNetV2 can effectively distinguish residential and commercial buildings from street-level imagery, thereby addressing the primary objective of this study, namely automated building-function classification.

The relatively symmetric distribution of false positives and false negatives indicates that the model does not strongly favor one class over the other. Balanced performance is particularly important in practical applications, where systematic bias toward either residential or commercial categories could lead to incorrect downstream decisions. The results suggest that the learned feature representations are sufficiently discriminative for both classes in the illustrated Figure 4.

True Label	Commercial	Residential
	87	7
Residential	7	75
	Commercial	Residential
	Predicted Label	

Figure 4. Confusion matrix for the performance of the model on the test data.

3.4 Analysis of Misclassification Cases

Although overall performance was high, some misclassifications were observed. Visual inspection of misclassified samples revealed that most errors occurred in ambiguous scenarios. Some residential buildings contained small commercial activities, such as home-based shops with minimal signage, making them visually similar to commercial structures. Conversely, certain commercial buildings with house-like façades and limited signage were sometimes predicted as residential.

These observations suggest that the model relies heavily on façade-level visual cues, including storefront structures, business signage, display windows, building entrances, and architectural design patterns. When such features are clearly visible, both residential and commercial buildings can be distinguished with high accuracy. However, prediction becomes more challenging when these visual indicators are absent, partially obscured, or shared between categories.

The misclassification patterns also reveal an inherent limitation of appearance-based building classification. In many urban environments, mixed-use buildings combine residential and commercial functions within the same structure. For example, small retail shops frequently operate on the ground floor of residential houses, creating visual characteristics that overlap between categories. Such ambiguity can make classification difficult even for human observers and suggests that certain prediction errors may arise from the complexity of the problem rather than model deficiencies alone.

Furthermore, the relatively balanced number of false positives and false negatives indicates that the model does not exhibit a strong bias toward either class. This suggests that MobileNetV2 successfully learned discriminative features from both residential and commercial buildings rather than overfitting to dominant visual patterns within a single category. Future studies may improve performance by incorporating additional contextual information, such as neighboring buildings, street context, or multiple street-view perspectives, which could provide complementary cues for distinguishing visually similar building types.

3.5 Generalization, Limitations, and Practical Implications

The strong performance of MobileNetV2 demonstrates that lightweight deep learning architectures can effectively classify building function from street-level imagery. The balanced precision and recall values obtained on the testing set suggest that the model is capable of learning meaningful visual characteristics associated with both residential and commercial buildings rather than relying on a single dominant class. Furthermore, the relatively low numbers of false positives and false negatives indicate reasonable generalization to unseen images within the dataset's distribution.

Nevertheless, several limitations should be acknowledged. The dataset was collected from publicly available street-level imagery and may not fully represent the diversity of building styles, urban layouts, and commercial practices across regions. Building appearance can vary considerably across cities and countries, potentially affecting model performance when applied to new geographic contexts. In addition, some buildings exhibit mixed residential and commercial functions, making visual classification inherently challenging even for human observers.

From a practical perspective, automated residential–commercial classification can support differentiated service delivery, small business mapping, and urban infrastructure planning. For instance, digital service providers and municipalities may use such models to identify areas with higher commercial density or to estimate the distribution of business activity without conducting extensive field surveys.

4. CONCLUSION

This study investigated the use of deep learning for classifying residential and commercial buildings based on street-level images. A dataset consisting of 882 images was collected and processed through several preprocessing stages, including cleaning, labeling, cropping, and resizing. The dataset was divided into training, validation, and testing sets using a 60:20:20 ratio. Two deep learning architectures were implemented and evaluated, namely a baseline CNN and MobileNetV2 with transfer learning.

The experimental results demonstrate that MobileNetV2 achieved superior performance compared to the baseline CNN across multiple evaluation metrics. The optimal configuration was obtained using a batch size of 32 and a learning rate of 0.0001. Under this configuration, the model achieved an accuracy of 92.05%, with precision, recall, and F1-score of 91.46% on the testing dataset. The confusion matrix analysis further indicated a relatively balanced classification performance, with a small number of false positive and false negative predictions. These results show that MobileNetV2 is capable of effectively extracting relevant visual features from street-level building images and generalizing well to unseen data.

Overall, the findings show that MobileNetV2 combined with transfer learning is effective for extracting meaningful visual features from street-level building images and can achieve reliable classification performance. The results indicate that lightweight deep learning architectures can provide efficient and accurate solutions for building classification tasks and may support practical applications such as urban analysis, digital infrastructure planning, and automated building identification systems. Despite the promising results, several opportunities remain for future research. First, larger and more geographically diverse datasets should be collected to evaluate model robustness across different urban environments. Second, future studies may investigate additional building categories, including industrial, institutional, and mixed-use buildings. Third, integrating contextual information from surrounding environments or multiple street-view perspectives may further improve classification accuracy.

ACKNOWLEDGMENTS

The authors would like to thank the Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Riau, for providing the facilities and academic support necessary to conduct this research.

REFERENCES

- [1] A. Ziaee and G. Suter, "Benchmarking methods for classifying space functions and access elements in multi-unit apartment buildings," *Autom. Constr.*, vol. 176, 2025, doi: 10.1016/j.autcon.2025.106237.
- [2] N. Lazar and K. Chithra, "Benchmarking critical criteria for assessing sustainability of residential buildings in tropical climate," *Journal of Building Engineering*, vol. 45, 2022, doi: 10.1016/j.jobbe.2021.103467.
- [3] L. Dahlström, F. Johari, and J. Widén, "Regional building energy modelling: A residential building stock model for a Swedish island," *Energy Build.*, vol. 346, 2025, doi: 10.1016/j.enbuild.2025.116178.
- [4] D. Sadhukhan et al., "Estimating surface temperature from thermal imagery of buildings for accurate thermal transmittance (U-value): A machine learning perspective," *Journal of Building Engineering*, vol. 32, 2020, doi: 10.1016/j.jobbe.2020.101637.
- [5] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," in *International Geoscience and Remote Sensing Symposium (IGARSS)*, Institute of Electrical and Electronics Engineers Inc., 2018, pp. 204–207. doi: 10.1109/IGARSS.2018.8519248.
- [6] Y. Yuan, X. Shi, and J. Gao, "Building extraction from remote sensing images with deep learning: A survey on vision techniques," *Computer Vision and Image Understanding*, vol. 251, 2025, doi: 10.1016/j.cviu.2024.104253.
- [7] T. K. Soni and P. Pujari, "Emerging deep learning approaches for urban satellite image analysis: a survey on classification, segmentation, and change detection," *Evol. Intell.*, vol. 18, no. 5, 2025, doi: 10.1007/s12065-025-01090-2.
- [8] K. Chen, Z. Zou, and Z. Shi, "Building extraction from remote sensing images with sparse token transformers," *Remote Sens. (Basel)*, vol. 13, no. 21, 2021, doi: 10.3390/rs13214441.
- [9] B. Saritürk, D. Z. Seker, O. Oztürk, and B. Bayram, "Performance evaluation of shallow and deep CNN architectures on building segmentation from high-resolution images," *Earth Sci. Inform.*, vol. 15, no. 3, pp. 1801–1823, 2022, doi: 10.1007/s12145-022-00840-5.
- [10] S. T. Seydi, H. Rastiveis, B. Kalantar, A. A. Halin, and N. Ueda, "BDD-Net: An End-to-End Multiscale Residual CNN for Earthquake-Induced Building Damage Detection," *Remote Sens. (Basel)*, vol. 14, no. 9, 2022, doi: 10.3390/rs14092214.
- [11] H. Xu, Q. Huang, H. Liao, G. Nong, and W. Wei, "MFFP-Net: Building Segmentation in Remote Sensing Images via Multi-Scale Feature Fusion and Foreground Perception Enhancement," *Remote Sens. (Basel)*, vol. 17, no. 11, 2025, doi: 10.3390/rs17111875.
- [12] K. Diaio, J. Zhu, G. Liu, and M. Li, "MDTrans: Multi-scale and dual-branch feature fusion network based on Swin Transformer for building extraction in remote sensing images," *IET Image Process.*, vol. 18, no. 11, pp. 2930–2942, 2024, doi: 10.1049/ipr2.13145.
- [13] Q. Yuan, H. Z. Mohd Shafri, A. H. Alias, and S. J. Bin Hashim, "Multiscale semantic feature optimization and fusion network for building extraction using high-resolution aerial images and lidar data," *Remote Sens. (Basel)*, vol. 13, no. 13, 2021, doi: 10.3390/rs13132473.
- [14] Q. Yuan and B. Xia, "Cross-level and multiscale CNN-Transformer network for automatic building extraction from remote sensing imagery," *Int. J. Remote Sens.*, vol. 45, no. 9, pp. 2893–2914, 2024, doi: 10.1080/01431161.2024.2339199.
- [15] J. Herman, R. Zewail, T. Ogawa, and S. Elsagheer, "A Lightweight Transfer Learning-Based Model for Building Classification in Aerial Imagery," in *2023 15th International Conference on Computer Research and Development, ICCRD 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 181–186. doi: 10.1109/ICCRD56364.2023.10080643.
- [16] D. S. Adams, T. Hauser, H. L. Yang, and P. Li, "Multivariate Testing of Sampling Techniques to Address Class Imbalance in Building Use Type Classification," in *GeoAI 2024 - Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, S. Gao, G. Mai, S. Newsam, L. Yang, D. Lunga, D. Zhu, B. Martins, and S. Arundel, Eds., Association for Computing Machinery, Inc, 2024, pp. 15–26. doi: 10.1145/3687123.3698282.
- [17] A. Karkehabadi and S. Sadeghmalakabadi, "Evaluating Deep Learning Models for Architectural Image Classification: A Case Study on the UC Davis Campus," in *2024 IEEE 8th International Conference on Information and Communication Technology, CICT 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/CICT64037.2024.10899686.

- [18] O. M. Mogaka, R. Zewail, K. Inoue, and M. S. Sayed, "TinyEmergencyNet: a hardware-friendly ultra-lightweight deep learning model for aerial scene image classification," *J. Real. Time. Image Process.*, vol. 21, no. 2, 2024, doi: 10.1007/s11554-024-01430-y.
- [19] M. Yang, "Efficiency and Adaptability of Deep Learning Architectures: A Performance Comparison of CNN, Transformer, and Hybrid Models," in *2025 5th International Conference on Computer Vision, Application and Algorithm, CVAA 2025*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 1–4. doi: 10.1109/CVAA66438.2025.11193320.
- [20] S. P. Ramalingam and V. Kumar, "Building usage prediction in complex urban scenes by fusing text and facade features from street view images using deep learning," *Build. Environ.*, vol. 267, 2025, doi: 10.1016/j.buildenv.2024.112174.
- [21] S. Wang, "Development of approach to an automated acquisition of static street view images using transformer architecture for analysis of Building characteristics," *Sci. Rep.*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-14786-3.
- [22] U. Kumar Lillhore et al., "A precise model for skin cancer diagnosis using hybrid U-Net and improved MobileNet-V3 with hyperparameters optimization," *Sci. Rep.*, vol. 14, no. 1, p. 4299, 2024, doi: 10.1038/s41598-024-54212-8.
- [23] D. Yang, X. Gao, Y. Yang, K. Guo, K. Han, and L. Xu, "Advances and Future Prospects in Building Extraction From High-Resolution Remote Sensing Images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 18, pp. 6994–7016, 2025, doi: 10.1109/JSTARS.2025.3538662.
- [24] A. Shoaib, M. Vadiveloo, and S. P. Lim, "Performance Comparison of Convolutional Neural Network Deep Learning Architectures for Remote Sensing Image Segmentation," in *Lecture Notes in Electrical Engineering*, V. Thiruchelvam, R. Alfred, I. Z.I.B.A., H. Haviluddin, and A. Baharum, Eds., Springer Science and Business Media Deutschland GmbH, 2024, pp. 199–211. doi: 10.1007/978-981-97-2977-7_12.
- [25] B. Kalantar, S. S. Ojogbane, S. T. Seydi, A. A. Halin, S. Mansor, and N. Ueda, "A Deep Learning Approach for Automated Building Outlines Extraction in Compact Urban Environments," in *International Geoscience and Remote Sensing Symposium (IGARSS)*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 7491–7494. doi: 10.1109/IGARSS46834.2022.9884697.
- [26] T. D. Akinosho et al., "Deep learning in the construction industry: A review of present status and future innovations," *Journal of Building Engineering*, vol. 32, 2020, doi: 10.1016/j.job.2020.101827.
- [27] Y. Li, J. Lu, Y. Meng, Y. Luo, and J. Ren, "Exploring Urban Spatial Quality Through Street View Imagery and Human Perception Analysis," *Buildings*, vol. 15, no. 17, p. 3116, Aug. 2025, doi: 10.3390/buildings15173116.